УДК 004.932.75’1

ОБЗОР СКРЫТОЙ МАРКОВСКОЙ МОДЕЛИ В РАСПОЗНАВАНИИ СИМВОЛОВ

Попов А. О.
Белорусский Государственный Университет Информатики и Радиоэлектроники, Минск, Беларусь, andrei.papou96@gmail.com

Реферат. Статья описывает скрытую марковскую модель и ее применение в распознавании текстовых символов. Рассмотрены теоретические аспекты данной модели ее использования, проблемы, которые могут возникнуть при создании скрытой марковской модели.

Abstract. Article describes Hidden Markov Model and its usage in text character recognition. Theoretical aspects and issues, that might occur during Hidden Markov Model creation, were also considered.

На сегодняшний день потребность в оцифровке текстов, то есть потребность в ПО, которое переводило бы данные с бумажного носителя в цифровой, возрастает все больше и больше. Оптическое распознавание символов может применяться для распознавания текста из любого мультимедиа, такого как изображение, аудио, видео [1]. Именно эта необходимость сподвигла меня на обзор различных техник распознавания символов, а именно использование скрытой марковской модели для таких случаев. Данная модель позволяет распознавать текст или символы с очень высоким шансом. [2]

Основой для скрытой марковской модели является марковская цепь, которая может быть описана как весовой конечный автомат, который содержит в себе конечный набор N состояний $S = \{s_1, s_2, s_3, s_4 \ldots\}$ и набор переходов между этими состояниями. Для каждого состояния существует вероятность π_i, что модель начнет с именно этой точки. Сумма вероятностей для всех состояний будет равна единице. Также на основе набора вероятностей для каждого состояния создается набор вероятностей переходов $A = \{a_{ij}\}$. Данный набор описывает вероятность перехода из состояния i в состояние j. На рисунке 1 изображена простая марковская цепь с тремя состояниями: s1, s2, s3. Вероятность перехода из состояния s_i в состояние s_j отображена на дуге, которая соединяет эти два состояния.

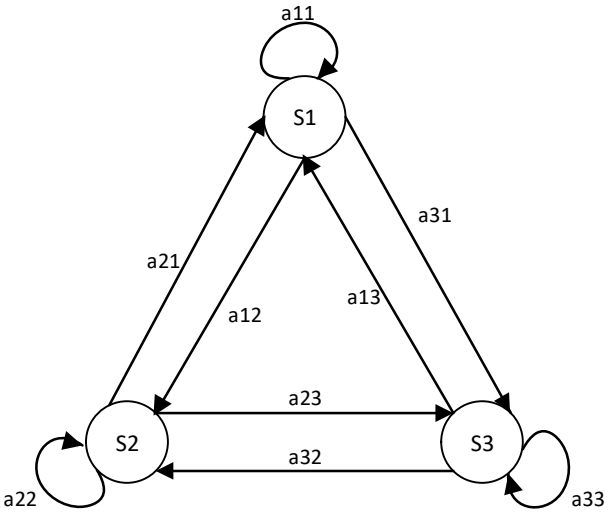

Рисунок 1 – Пример марковской модели с тремя состояниями

Марковская модель предполагает, что был создан марковским источником информации, в котором новые символы зависят только от фиксированного количества предшествующих

элементов. Данное количество зависит от порядка модели. Чаще всего применяются марковские модели первого и второго порядка в связи с тем, что при повышении порядка возрастает и сложность модели, что в свою очередь снижает ее полезность. Для марковской модели первого порядка вероятность определенного состояния зависит только от предыдущего. Определение имеет следующий вид.

$$P(q_t|q_{t-1} \dots q_1) = P(q_t|q_{t-1})$$

Для марковской модели второго порядка вероятность наступления состояния s_i в момент времени t зависит от состояния s_j, находящегося в моментах *t-2* и *t-1*. Количество m предыдущих состояний, от которых будет зависеть вероятность наступления следующего состояния, является порядком марковской модели.

Процесс описанный выше является наблюдаемым. Это означает, что все события, которые присутствуют в данной модели являются воспроизводимыми и физическими событиями. При распознавании текста не всегда есть возможность точно предсказать точную последовательность событий, которая приведет к желаемому результату. В этом помогает скрытая марковская модель. Данная модель представляет из себя дважды стохастический вариант марковской модели, в которой наблюдение за одним стохастическим процессом осуществляется при помощи набора других стохастических процессов, результатом которых является набор наблюдаемых символов. Скрытая марковская модель может быть представлена как взаимосвязанный набор состояний, которые соединены между собой набором вероятностей переходов. [5] В начале процесс инициализируется в каком-нибудь из состояний, после чего переходит в новое в зависимости от вероятности перехода. Далее по мере переходов в новое состояние вырабатывается набор символов, какой именно символ будет выбран из этой последовательности будет определен вероятностью результата, которая свойственна состоянию. В результате работы модели производится набор последовательность символов, а так как последовательностей событий, которые могут привести к определенной выходной последовательности символов, множество, последовательность состояний является скрытой.

Слегка изменим модель марковского процесса, представленную на рисунке 1, преобразовав ее в скрытую марковскую модель. Допустим 0 и 1 – это наблюдаемые символы каждого состояния. Тогда b_{ij} отвечает за вероятность результата каждого состояния. Пример данной модели представлен ниже на рисунке 2.

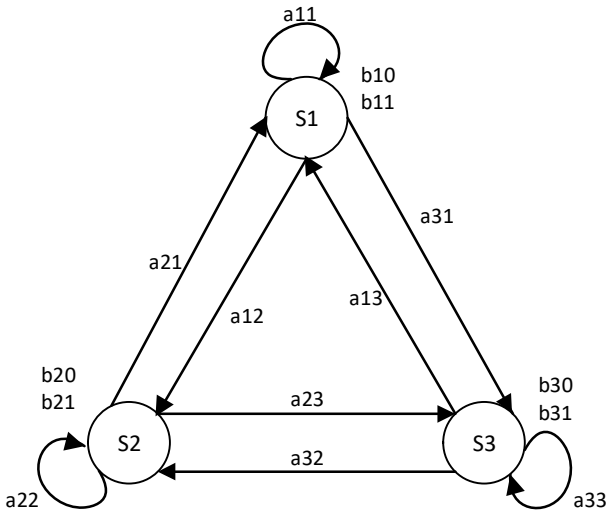

Рисунок 2 – Пример скрытой марковской модели с тремя состояниями

Каждое состояние содержит набор параметров, где $S = \{ s_1, s_2, s_3, s_4 \dots \}$ – конечный набор N состояний, V = {υ_1, υ_2, υ_3, υ_4 … m} – набор размерностью M возможных символов в словаре,

вероятность π_i, что модель начнет с именно этой точки, набор вероятностей переходов перехода из состояния i в состояние j $A = \{a_{ij}\}$ и $B = \{b_i(\upsilon_k)\}$ – набор вероятностей результата, где $b_i(\upsilon_k)$ вероятность генерации символа υ_k в состоянии i.

Однако скрытая марковская модель тоже имеет свои недостатки. [3] Первая проблема – это оценка. Предположим, что дана последовательность наблюдений $O = \{o_1, o_2, o_3, o_4 \dots o_t\}$ и модель λ. Как высчитать вероятность того, что последовательность наблюдений была выработана этой моделью? Данную проблему решает алгоритм прямого-обратного хода.

Вторая проблема – это декодирование. Много различных последовательностей состояний могут выдать одну и ту же наблюдаемую последовательность. Для каждой наблюдаемой последовательности нужно выбрать те состояния, которые будут иметь наивысшую вероятность выработки наблюдений для нахождения оптимальной последовательности символа. Для решения такой задачи часто используется алгоритм Витерби.

Последней проблемой является обучение. Задача состоит в правильном подборе параметров модели для того, чтобы как можно больше увеличить возможность генерации наблюдаемой последовательности. Наиболее популярные методы по обучению модели являются метод максимального правдоподобия и алгоритм Баума-Велша, который в свою очередь использует алгоритм прямого-обратного хода. Основные ошибки, которые могут встречаться при обучении модели по распознаванию символов – это пропуск символа, вставка лишних символов, замена одного символа другим, замена двух символов на один и наоборот и замена двух символов на два других символа. [4] Примеры данных ошибок представлены на таблице 1.

Таблица 1. – Типичные ошибки при распознавании символов.

Тип ошибки	Пример ошибки
Пропуск символа	deer → dee
Вставка символа	cat → c at
Замена одного символа другим	r → t; e → c; a →n
Замена двух символов одним	ni → m; ii → u; tl → k
Замена одного символа двумя	n → ii; u → ii; m → ni
Замена двух символов двумя другими символами	rm → nn; rw → nr

Также стоит отметить, что обучение и тренировка модели являются двумя важными шагами в классификации.[6] Схема, показывающая все шаги в классификации представлена ниже на рисунке 3.

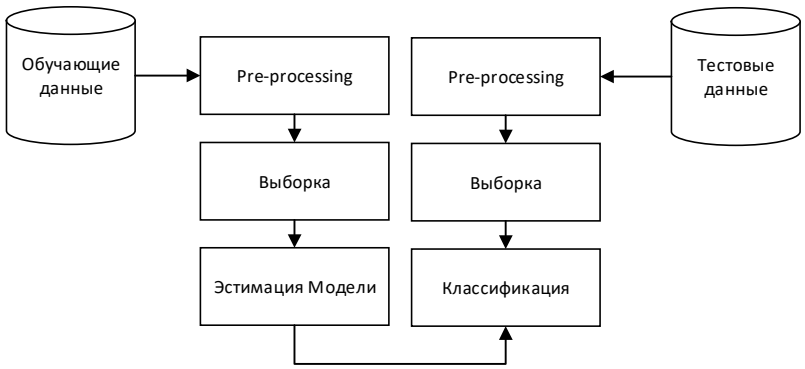

Рисунок 3 – Шаблон классификационного процесса

Изначально данные приводятся к виду, подходящему для обучения, далее происходит выборка, цель которой заключается в уменьшении количества данных, путем взятия только нужной информации. При эстимации модели из конечного набора векторов производится оценка модели для каждого класса в наборе данных.

Список использованной литературы

1. Hiral Modi, M. C. Parikh “A Review on Optical Character Recognition Techniques” International Journal of Computer Applications (0975 – 8887) Volume 160 – No 6, February 2017

2. J. Hu, S. G. Lim, and M. K. Brown, “Writer independent on-line handwriting recognition using an HMM approach,” J. PATTERN Recognit. Soc., vol. 33, стр. 133– 147, 2000

3. L.R Rabiner and B.H Juang, “An Introduction to Hidden Markov Model,” ASSP, vol. 3, no. 1, стр. 4–16, 1986.

4. J. Esakov, D. P. Lopresti, and J. S. Sandberg, “Classification and distribution of optical character recognition errors,” in Proceedings of the IS&T/SPIE International Symposium on Electronic Imaging, (San Jose, CA), February 1994.

5. S. M. Thede and M. P. Harper, “A second-order Hidden Markov Model for part-of-speech tagging,” in Proceedings of the 37th Annual Meeting of the ACL, pp. 175–182, 1999.

6. Saish Bhende, Kutub Thakur, Jason Teseng, Md Liakat Ali, Nan Wang “Character Recognition Using Hidden Markov Models”